A tutorial on spatio-temporal disease risk modelling in R using Markov chain Monte Carlo simulation and the CARBayesST package

Duncan Lee[a,*]

[a]*School of Mathematics and Statistics, University of Glasgow*

Abstract

Population-level disease risk varies in space and time, and is typically estimated using aggregated disease count data relating to a set of non-overlapping areal units for multiple consecutive time periods. A large research base of statistical models and corresponding software has been developed for such data, with most analyses being undertaken in a Bayesian setting using either Markov chain Monte Carlo (MCMC) simulation or integrated nested Laplace approximations (INLA). This paper presents a tutorial for undertaking spatio-temporal disease modelling using MCMC simulation, utilising the CARBayesST package in the R software environment. The tutorial describes the complete modelling journey, starting with data input, wrangling and visualisation, before focusing on model fitting, model assessment and results presentation. It is illustrated by a new case study of pneumonia mortality at the local authority level in England, and answer important public health questions including the effect of covariate risk factors, spatio-temporal trends, and health inequalities.

Keywords: Bayesian inference, CARBayesST, Spatio-temporal modelling

1. Introduction

Population-level disease risk varies in space and time between sub-populations, due to variation in environmental exposures and the prevalence of risk inducing behaviours such as

*Corresponding author - Duncan Lee, School of Mathematics and Statistics, University of Glasgow, Glasgow, G12 8SQ

Email address: Duncan.Lee@glasgow.ac.uk (Duncan Lee)

smoking. Estimating the spatio-temporal variation in disease risk is an important endeavour for researchers and policy-makers alike, because it allows them to answer important public health questions such as: (i) where are the highest-risk areas for disease that can be targeted for an intervention; (ii) is disease risk increasing or decreasing over time; (iii) how big are the health inequalities, and are they changing over time; and (iv) what environmental or social factors affect the risk of disease. In the above, health inequalities are the differences in disease risk between different communities (World Health Organisation, 2013) and are often driven by poverty, with poorer populations typically exhibiting higher disease risks than more affluent populations.

Data summarising population level disease incidence are commonly used to answer these questions, which have been spatio-temporally aggregated to a set of K non-overlapping areal units for N consecutive time intervals. These data thus comprise a $K \times N$ matrix of spatio-temporal observed disease counts, which are augmented by matrices of expected counts computed using indirect standardisation to allow for varying population demographics, and covariate risk factors. A large number of models have been developed for estimating the spatio-temporal variation in disease risk, and comprehensive reviews are given by Lawson and Lee (2017) and Lawson (2018). The class of conditional autoregressive (CAR, Besag et al., 1991) models are commonly used to represent the spatially correlated variation in disease risk, and are used as a prior distribution for a set of spatially structured random effects. Spatio-temporal extensions of CAR models have been proposed by numerous authors, including Bernardinelli et al. (1995), Knorr-Held (2000) and Rushworth et al. (2014). These models assume that disease risk varies smoothly in space and time, and thus account for the inherent spatio-temporal autocorrelation typically observed amongst the disease data. A Bayesian approach to inference is typically adopted, using either Markov chain Monte Carlo (MCMC, Robert and Casella, 2010) simulation or Integrated Nested Laplace Approximations (INLA, Rue et al., 2009)).

A number of software packages have been developed to allow researchers to fit spatio-temporal conditional autoregressive type models in a Bayesian setting, including `WinBUGS` (Lunn et al., 2000) and `STAN` (Morris et al., 2019). However, these packages can be dif-

ficult to use for novice users, and also lack the ability to undertake additional analyses, such as data visualisation and wrangling, within the same software environment. Therefore the R software environment (https://cran.r-project.org) has a number of packages for spatio-temporal areal unit modelling, including the INLA package (Martins et al., 2013) using integrated nested Laplace approximations, and CARBayes (spatial modelling, Lee, 2013) and CARBayesST (spatio-temporal modelling, Lee et al., 2018) both using MCMC simulation. Both INLA and CARBayes / CARBayesST can fit a range of different spatio-temporal disease risk models for areal unit data, while INLA can also model geostatistical (Lindgren et al., 2011) and point process (Illian et al., 2012) spatio-temporal data. The main advantage of INLA is computational speed, because it uses Laplace approximations to estimate the posterior distribution of the unknown parameters, rather than drawing repeated samples from the posterior distribution as MCMC does. In contrast, while slower than INLA, CARBayesST is potentially easier to use for novice users, because: (i) all models can be implemented via a simple one-line function call requiring the user to specify few arguments for a default analysis, even when specifying the spatio-temporal data structures; and (ii) important epidemiological quantities such as posterior exceedance probabilities are straightforward to produce because you have direct access to samples from the posterior risk distribution. Additionally, CARBayeST allows users to fit localised spatial autocorrelation models such as that proposed by Rushworth et al. (2017), which is not possible using the INLA package. These localised correlation models recognise that disease risk in geographically adjacent areal units may not always be similar (correlated), which stems from the seminal work of Womble (1951).

Excellent tutorials for how to use the INLA package for spatio-temporal modelling have been written by Blangiardo et al. (2013) and Moraga (2018), but no such tutorial exists for spatio-temporal modelling with CARBayesST using MCMC simulation. Therefore this paper fills that gap, and presents a tutorial describing how to undertake spatio-temporal Bayesian areal unit modelling via MCMC simulation using the CARBayesST package. Lee et al. (2018) provides a general vignette for the CARBayesST software including a description of the suite of models that the package can fit, where as here we provide a specific tutorial on spatio-temporal disease risk modelling aimed at applied public health researchers. We also note that the purely spatial modelling package CARBayes has an identical syntax to CARBayesST,

and hence will be straightforward to use following a similar approach to that described here.

This tutorial is illustrated by a new study of pneumonia mortality in England at the local authority level. The data and the questions motivating the analysis are described in Section 2, while exploratory analysis is presented in Section 3. Spatio-temporal model fitting and model assessment is outlined in Section 4, while the results of the analysis are presented in Section 5. Finally, this tutorial finishes with an outline of future developments in Section 6. The data, shapefiles and `R` code used in this tutorial are available to download from https://github.com/duncanplee/Spatio-temporal-modelling-tutorials, which also includes a video tutorial illustrating the analysis presented here.

2. Motivating study

The study region is mainland England, which has a population of around 55 million people and has been partitioned into $K = 322$ local authorities. Data are available at yearly intervals between 2002 to 2017 inclusive, yielding $N = 16$ time periods. The disease data $\{\mathtt{Y}_{kt}\}$ are counts of the numbers of pneumonia mortalities (International Classification of disease ICD-10 codes J12-J18) in local authority k during year t, and were obtained from NHS digital https://digital.nhs.uk/. These counts vary between 10 and 680 in each local authority and year, with a median value of 75. The local authorities have different population sizes and demographic structures, and indirect standardisation is used to account for the fact that areas with larger and more elderly populations are likely to exhibit more mortalities. Specifically, the population in local authority k during year t is split into R strata based on age and sex (e.g. females 0-4, females 5-9, etc.), and n_{ktr} denotes the number of people in strata r. These strata specific population sizes n_{ktr} are multiplied by national strata specific pneumonia mortality rates γ_r averaged over the study period 2002-2107, and the results are summed over strata to give the expected number of pneumonia mortalities in local authority k during year t. Mathematically these expected counts are computed by

$$\mathtt{E}_{kt} = \sum_{r=1}^{R} n_{ktr}\gamma_r, \tag{1}$$

and represent the number of mortalities expected if national age and sex specific pneumonia mortality rates averaged over 2002-2017 applied to local authority k during year t. Additionally, we also have two covariates that may explain the spatio-temporal variation in pneumonia mortality risk, the first of which is a proxy measure of socio-economic deprivation (poverty). Socio-economic deprivation is well known to affect morbidity and mortality, and is one of the main drivers of health inequalities (World Health Organisation, 2013). In this study we have access to the English Index of multiple deprivation (IMD, Department for Communities and Local Government, 2015) in 2015, which for local authority k is denoted by $\texttt{IMD}_k$. The IMD is a composite index of deprivation comprising data from 7 different domains, including: barriers to housing, crime, education, employment, health, income and living environment. As the calculation of the index changes from one year to the next, we only use data for one year, meaning this covariate will only capture the spatial variation in disease risk.

The other covariate we have is a measure of fine particulate matter air pollution known as $\text{PM}_{2.5}$ (measured in μgm^{-3}), which is a mixture of solid and liquid particles in the air that are less than 2.5 micrometres in diameter. Measured concentrations of $\text{PM}_{2.5}$ are available from https://uk-air.defra.gov.uk, but the network of monitors is not dense at the local authority level required for this study. Therefore instead we utilise annual average modelled concentrations from the Pollution Climate Mapping (PCM) model developed for the Department for the Environment, Food and Rural Affairs (DEFRA), which are freely available from https://uk-air.defra.gov.uk/data/pcm-data. The model estimates annual average $\text{PM}_{2.5}$ concentrations on a $1km^2$ square grid, which need to be spatially realigned with the irregularly shaped local authorities. This is achieved by computing the average concentration in each local authority by

$$\texttt{PM25}_{kt} = \frac{1}{q_k} \sum_{i|\mathbf{g}_i \in \mathcal{A}_k} \texttt{PM25}_{it}, \tag{2}$$

where the average is taken over all grid squares whose centroid $\mathbf{g}_i$ lies within local authority $\mathcal{A}_k$. In the above equation q_k denotes the number of grid square centroids that lie within the kth local authority $\mathcal{A}_k$, while $\texttt{PM25}_{it}$ denotes the annual average concentration of $\text{PM}_{2.5}$

for year t and grid square with centroid $\mathbf{g}_i$. Our aim in analysing these data is to answer 3 key questions of public health importance, which are:

1. What effects do the air pollution and socio-economic covariates have on disease risk?
2. What is the temporal trend in disease risk and which local authorities exhibit the highest risks of disease?
3. How big are the health inequalities in pneumonia mortality risk across England, and are these inequalities increasing or decreasing over time?

3. Exploratory analysis

3.1. Reading in and visualising the spatio-temporal trends in the data

The disease and covariate data are stored in `EnglandLUAdata.csv`, which has a unique local authority code (`Code`) and year (`Year`) combination for each row of the data set. The local authority code is accompanied by the name of local authority (`Name`), and the remaining 4 variables were described in the previous section. The data can be read into `R` and the variable names visualised using the following commands.

```
dat <- read.csv(file="EnglandLUAdata.csv")
head(dat, n=3)
```

```
##   Code                    Name Year   Y        E     PM25    IMD
## 1 00AB Barking and Dagenham LB 2002 145 78.04092 11.23563 34.635
## 2 00AB Barking and Dagenham LB 2003 210 75.89981 16.92847 34.635
## 3 00AB Barking and Dagenham LB 2004 130 75.55028 16.70258 34.635
```

Note, it is simplest in practice to put the data in the same folder as the `R` script file you are writing your analysis in, and then in `Rstudio` setting the working directory to the location of the source file. The first step in an exploratory analysis of these data is to compute the standardised mortality ratio (SMR) for each local authority and year combination to measure the mortality risk, which is computed by dividing the observed number of mortalities (`Y`) by the expected number of mortalities (`E`). It is added to the data set using the `dplyr` package via the following code.

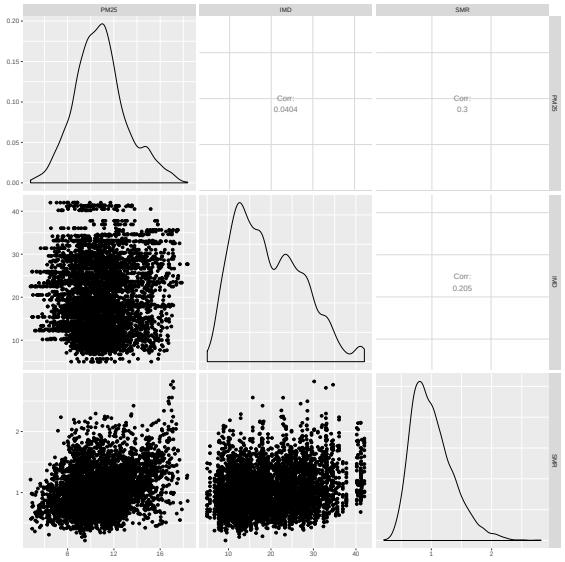

Figure 1: Scatterplots displaying the relationships between each covariate and the SMR.

```
library(dplyr)
dat <- dat %>% mutate(SMR=dat$Y/dat$E)
```

If the SMR equals 1 the observed and expected mortality counts are equal which represents an average risk area relative to the entire study region, while an SMR greater than 1 denotes a high risk area. For example, an SMR of 1.2 corresponds to a 20% increased risk compared to the English national average over the 16 year study period. Similarly an SMR of 0.9 corresponds to a 10% decreased risk compared to the national average. The relationship between the SMR and the two covariates can be visualised via scatterplots (Figure 1) using the `GGally` library via the code below, which gives an exploratory answer to our first motivating question.

```
library(GGally)
ggpairs(dat, columns=6:8)
```

The figure shows there are low to medium correlations between the SMR and the two covariates, suggesting that the latter might be significant predictors of disease risk. Visualising the spatio-temporal trends in the SMR will give an initial answer to the second and third

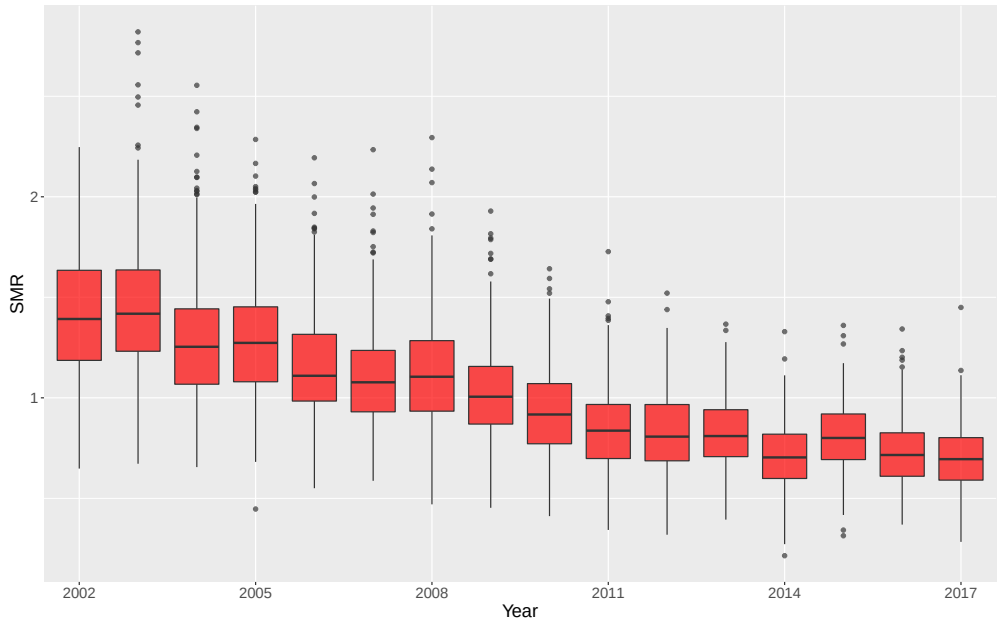

Figure 2: Boxplots displaying the temporal trend in the SMR.

motivating question for this analysis, and the temporal trend in disease risk can be visualised via boxplots of the SMR for each year (Figure 2) using the code below.

```
library(ggplot2)
ggplot(dat, aes(x = factor(Year), y = SMR)) +
        geom_boxplot(fill="red", alpha=0.7) +
        scale_x_discrete(name = "Year", breaks=c(2002, 2005, 2008, 2011, 2014, 2017),
                labels=c("2002", "2005", "2008", "2011", "2014", "2017")) +
        scale_y_continuous(name = "SMR") +
        theme(text=element_text(size=16), plot.title=element_text(size=18, face="bold"))
```

The figure shows that the SMR appears to decrease substantially over time, and is lowest in the final year of 2017 compared to any of the preceding years. The magnitude of the health inequalities also appears to have decreased, as there is markedly less variation in disease risk (the boxplots are narrower) in the later years. To view the average spatial pattern in the SMR we first need to read the shapefiles containing the local authority boundaries into `R`, which can be done using functionality of the `rgdal` package as shown below.

```
library(rgdal)
LA <- readOGR(dsn = "LocalAuthorities.shp")
```

Then the average SMR for each local authority can be computed via the code below.

```
by_LA <- group_by(dat, Code)
averageSMR <- summarize(by_LA, SMR = mean(SMR, na.rm=T))
```

This new data set `averageSMR` then needs to be combined with the local authority boundaries via

```
averageSMR.LA <- merge(x=LA, y=averageSMR, by.x="lad09cd", by.y="Code", all.x=FALSE)
```

which gives a `SpatialPolygonsDataFrame` object. The SMR variable can then be superimposed onto an OpenStreetMap using functionality from the `leaflet` package, and the resulting map is displayed in Figure 3.

```
library(leaflet)
averageSMR.LA <- merge(x=LA, y=averageSMR, by.x="lad09cd", by.y="Code", all.x=FALSE)
averageSMR.LA.ll <- spTransform(averageSMR.LA, CRS("+proj=longlat +datum=WGS84 +no_defs"))
variable <- averageSMR.LA.ll@data$SMR
colours <- colorNumeric(palette = "YlOrBr", domain = variable, reverse=FALSE)
leaflet(data=averageSMR.LA.ll) %>%
        addTiles() %>%
        addPolygons(fillColor = ~colours(variable), color="", fillOpacity = 0.7,
                weight = 1, smoothFactor = 0.5, opacity = 1.0) %>%
        addLegend(pal = colours, values = variable, opacity = 1, title="SMR") %>%
        addScaleBar(position="bottomleft")
```

The second line of the code above transforms the coordinate reference system of the `averageSMR.LA` object to longitude and latitude to be compatible with OpenStreetMap. The map shows that the main clusters of high risk local authorities are around the north west urban communities of Liverpool, Manchester and Sheffield, while lower risk areas are generally rural

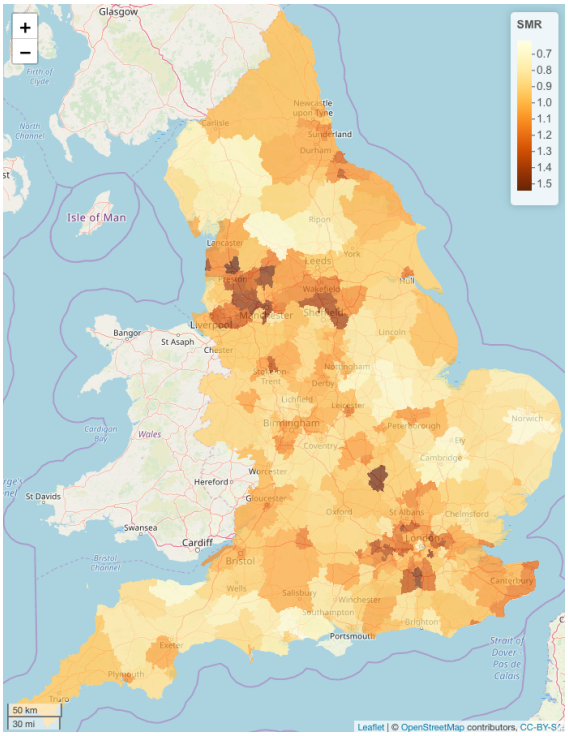

Figure 3: Spatial map of the average SMR between 2002 and 2017.

3.2. Naive regression modelling ignoring spatio-temporal autocorrelation

Before incorporating spatio-temporally autocorrelated random effects into the model, one should check whether the available covariates capture all of the spatio-temporal autocorrelation in the disease data. Thus a simple initial model is a Poisson log-linear model, which for our data is given by

$$\begin{aligned} Y_{kt} &\sim \text{Poisson}(E_{kt}\theta_{kt}) \quad \text{for } k = 1, \ldots, K, \ t = 1, \ldots, N \\ \ln(\theta_{kt}) &= \beta_0 + \beta_1 \texttt{IMD}_k + \beta_2 \texttt{PM25}_{kt}, \end{aligned} \tag{3}$$

where θ_{kt} is the risk of disease in local authority k during year t relative to the expected counts E_{kt}, and is on the same scale as the SMR. This model can be fitted using maximum likelihood estimation via the `glm()` function, or in a Bayesian setting using MCMC simulation using the function `S.glm()` within the `CARBayes` package. Here we adopt the former approach for simplicity, and fit the above model as shown below:

```
model1 <- glm(formula=Y~offset(log(E)) + IMD + PM25, family="poisson", data=dat)
round(cbind(model1$coefficients, confint(model1)),4)
```

```
##                       2.5 %  97.5 %
## (Intercept) -0.6529 -0.6681 -0.6377
## IMD          0.0082  0.0079  0.0086
## PM25         0.0443  0.0430  0.0455
```

From the point estimates and 95% confidence intervals both covariates initially exhibit significant effects on pneumonia mortality risk, with increasing levels of socio-economic deprivation and air pollution being associated with an increased risk. To check for the presence of spatial autocorrelation we first add the residuals from this simple model to the data set and then extract the residuals for a single year. We illustrate this process for 2010, where the last line transforms the residuals into a `SpatialPolygonsDataFrame` object.

```
dat$residuals <- residuals(model1)
residuals2010 <- filter(dat, Year==2010)
residuals2010.LA <- merge(x=LA, y=residuals2010, by.x="lad09cd", by.y="Code", all.x=FALSE)
```

A commonly used measure of spatial autocorrelation is Moran's I statistic (Moran, 1950), which is a spatially altered version of Pearson's correlation coefficient. It is given by

$$I = \frac{K \sum_{k=1}^{K} \sum_{j=1}^{K} w_{kj}(r_k - \bar{r})(r_j - \bar{r})}{\left(\sum_{k=1}^{K} \sum_{j=1}^{K} w_{kj}\right) \sum_{k=1}^{K} (r_k - \bar{r})^2}, \tag{4}$$

where r_k denotes the residual for the kth local authority. Additionally, $\mathbf{W} = (w_{kj})$ is a $K \times K$ neighbourhood or adjacency matrix, which denotes whether each pair of local authorities are close together. Typically a binary specification is taken, where $w_{kj} = 1$ if local authorities (k, j) share a common border, and $w_{kj} = 0$ otherwise. Other specifications are possible, and a review is given by Earnest et al. (2007). The value of Moran's I statistic lies between (-1,1), with a value of zero corresponding to spatial independence while a value of 1 denotes strong positive spatial autocorrelation. Before computing this statistic the neighbourhood matrix $\mathbf{W}$ needs to be constructed using the following code.

```
library(spdep)
W.nb <- poly2nb(residuals2010.LA, row.names = residuals2010.LA@data$lad09cd)
W <- nb2mat(W.nb, style = "B")
W.list <- nb2listw(W.nb, style = "B")
```

The neighbourhood matrix is the `W` object, while the `W.list` object is a list type variant of this object required in the computation of Moran's I. Further details on the construction of these objects and other alternatives are available from Bivand et al. (2013) and the `spdep` package. A permutation test with the null hypothesis of spatial independence can be conducted based on Moran's I statistic using the following code, where the p-value is based on 10,000 random permutations of the data.

```
moran.mc(x = residuals2010.LA$residuals, listw = W.list, nsim = 10000)
```

```
##  Monte-Carlo simulation of Moran I
##
## data:  residuals2017.LA$residuals
## weights: W.list
## number of simulations + 1: 10001
##
## statistic =0.27104, observed rank = 10001, p-value = 9.999e-05
## alternative hypothesis: greater
```

The test shows strong residual spatial autocorrelation, with a Moran's I value of 0.271 and a p-value much less than 0.05. One could now assess the presence of residual temporal autocorrelation using the `acf()` function, but with only $N = 16$ time periods the results would not be that reliable and hence we don't do that here.

4. Bayesian spatio-temporal modelling using MCMC simulation

As summarised in the introduction, a large number of models have been proposed for estimating the spatio-temporal trends in disease risk, and as our disease outcome variable is a count, they have the general form

$$
\begin{aligned}
Y_{kt} &\sim \text{Poisson}(E_{kt}\theta_{kt}) \quad \text{for } k = 1, \ldots, K,\ t = 1, \ldots, N \\
\ln(\theta_{kt}) &= \beta_0 + \beta_1 \texttt{IMD}_k + \beta_2 \texttt{PM25}_{kt} + \psi_{kt},
\end{aligned} \tag{5}
$$

where ψ_{kt} is the random effect for local authority k and time period t. The regression parameters are typically assigned independent and weakly informative normal priors, and the default prior specification in `CARBayesST` is $\beta_j \sim \text{N}(0, 100000)$. While independent normal priors are enforced for each β_j in the software, the prior mean and variance can be changed using the arguments `prior.mean.beta, prior.var.beta` in the function call when fitting the model, see the individual helpfiles for more details.

The spatio-temporal structure of the random effects $\{\psi_{kt}\}$ depends on the goal of the analysis, and 3 commonly used models are outlined below.

- **Correlated linear time trends** - $\psi_{kt} = \beta_1 + \phi_k + (\alpha + \delta_k)(t - \bar{t})/N$, where $\bar{t} = (1/N)\sum_{t=1}^{N} t$, so that the time trend $(t - \bar{t})/N$ runs over a centred unit interval. Additionally, $\{\phi_k\}$ and $\{\delta_k\}$ are sets of spatially autocorrelated random effects, which allow the linear temporal trends to have spatially varying intercepts and slopes. This model was originally proposed by Bernardinelli et al. (1995), and is available in `CARBayesST` via the `ST.CARlinear()` function.

- **Spatio-temporal main effects and an interaction** - $\psi_{kt} = \phi_k + \delta_t + \gamma_{kt}$, where $\{\phi_k\}$ and $\{\delta_t\}$ are respectively sets of spatially and temporally autocorrelated random effects, while $\{\gamma_{kt}\}$ are space-time interactions. This model was originally proposed by Knorr-Held (2000), and is available in `CARBayesST` via the `ST.CARanova()` function.

- **Spatially autocorrelated first-order autoregressive process** - $\boldsymbol{\psi}_t = \rho_T \boldsymbol{\psi}_{t-1} + \boldsymbol{\epsilon}_t$, where $\boldsymbol{\psi}_t = (\psi_{1t}, \ldots, \psi_{Kt})$ denotes the vector of random effects for all areal units at time t, and the vector of errors $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{Kt})$ is modelled as spatially autocorrelated. This model was used by Rushworth et al. (2014), and is available in `CARBayesST` via the `ST.CARar()` function.

In this tutorial the aim is to quantify the evolution of the spatial pattern in disease risk over time, so we use the spatially autocorrelated first-order autoregressive process model given by $\boldsymbol{\psi}_t = \rho_T \boldsymbol{\psi}_{t-1} + \boldsymbol{\epsilon}_t$. Temporal autocorrelation is controlled by the mean function $\rho_T \boldsymbol{\psi}_{t-1}$, while spatial autocorrelation is controlled by the covariance structure of $\boldsymbol{\epsilon}_t$. The latter is modelled as spatially autocorrelated and given by $\boldsymbol{\epsilon}_t \sim \mathrm{N}(\mathbf{0}, \tau^2 \mathbf{Q}(\mathbf{W}, \rho_S)^{-1})$, where τ^2 is the process variance. The precision matrix is given by $\mathbf{Q}(\mathbf{W}, \rho_S) = \rho_S(\mathrm{diag}[\mathbf{W1}] - \mathbf{W}) + (1 - \rho_S)\mathbf{I}$, where $(\mathbf{1}, \mathbf{I})$ are a $K \times 1$ vector of ones and the $K \times K$ identity matrix respectively Spatial autocorrelation is induced by the neighbourhood matrix $\mathbf{W}$ defined above, and if $w_{kj} = 1$ then the random errors $(\epsilon_{kt}, \epsilon_{jt})$ are modelled as spatially autocorrelated, while if $w_{kj} = 0$ then $(\epsilon_{kt}, \epsilon_{jt})$ are assumed to be conditionally independent. Thus (ρ_S, ρ_T) respectively control the levels of spatial and temporal autocorrelation, with values of 0 corresponding to independence while a value of 1 corresponds to strong autocorrelation. The precision matrix $\mathbf{Q}(\mathbf{W}, \rho_S)$ corresponds to the conditional autoregressive (CAR) prior proposed by Leroux et al. (2000), and this multivariate specification ($\boldsymbol{\epsilon}_t \sim \mathrm{N}(\mathbf{0}, \tau^2 \mathbf{Q}(\mathbf{W}, \rho_S)^{-1})$) is equivalent to

$$\begin{aligned} \epsilon_{kt} | \boldsymbol{\epsilon}_{-kt}, \mathbf{W} &\sim \mathrm{N}\left(\frac{\rho_S \sum_{j=1}^{K} w_{kj} \epsilon_{jt}}{\rho_S \sum_{j=1}^{K} w_{kj} + 1 - \rho_S}, \frac{\tau^2}{\rho_S \sum_{j=1}^{K} w_{kj} + 1 - \rho_S} \right), \\ \tau^2 &\sim \text{Inverse-Gamma}(1, 0.01), \\ \rho_S, \rho_T &\sim \text{Uniform}(0, 1), \end{aligned} \tag{6}$$

where the last 2 lines give weakly informative priors for the remaining parameters and $\boldsymbol{\epsilon}_{-kt} = (\epsilon_{1t}, \ldots, \epsilon_{k-1t}, \epsilon_{k+1t}, \ldots, \epsilon_{Kt})$. Whilst the default prior specification for τ^2 is $\tau^2 \sim$ Inverse-Gamma$(1, 0.01)$, the hyperparameters $(1, 0.01)$ can be changed by the user via the `prior.tau2` argument in the function call when fitting the model, see the helpfile for `ST.CARar()` for more details. If $\rho_S = 1$ the model simplifies to the intrinsic CAR prior proposed Besag et al. (1991), while if $\rho_S = 0$ the errors ϵ_{kt} are independent with mean zero and a constant variance τ^2, i.e. $\epsilon_{kt} \sim \mathrm{N}(0, \tau^2)$.

Before fitting this model one must re-order the data in `dat`, because the software `CARBayesST` requires the data to be ordered so that the first K data points relate to all the spatial units for time period 1, the next K data points relate to all the spatial units for time period 2

and so on. Additionally, the spatial units must be ordered in the same way as specified by the neighbourhood matrix `W`. These ordering constraints can be implemented using the code below, which first creates a spatial ordering identifier (`spatialorder`) and then arranges the data by that identifier.

```
lookup <- data.frame(Code=residuals2010.LA@data$lad09cd,
                spatialorder=1:nrow(residuals2010.LA@data))
dat.temp <- merge(x=dat, y=lookup, by="Code")
dat.ordered <- arrange(dat.temp, Year, spatialorder)
```

The model can then be fitted to the ordered data using MCMC simulation via the `ST.CARar()` function using the code below.

```
library(CARBayesST)
chain1 <- ST.CARar(formula=Y~offset(log(E)) + PM25 + IMD, family="poisson",
        data=dat.ordered, W=W, burnin=200000, n.sample=2200000, thin=1000, verbose=FALSE)
chain2 <- ST.CARar(formula=Y~offset(log(E)) + PM25 + IMD, family="poisson",
        data=dat.ordered, W=W, burnin=200000, n.sample=2200000, thin=1000, verbose=FALSE)
chain3 <- ST.CARar(formula=Y~offset(log(E)) + PM25 + IMD, family="poisson",
        data=dat.ordered, W=W, burnin=200000, n.sample=2200000, thin=1000, verbose=FALSE)
```

The above code runs the model 3 times to generate MCMC samples from 3 independent Markov chains. Each chain is run for 2,200,000 samples (`n.sample`), of which 200,000 are removed as the burnin period (`burnin`) and the remaining 2,000,000 samples are thinned by 1,000 (`thin`) to remove almost all of the correlation amongst the samples. This leaves 6,000 samples for inference overall, with 2,000 coming from each chain. The first step is to assess whether the Markov chains have converged, which can be done in numerous ways. The simplest check is to draw a *traceplot* of the samples for each parameter, and convergence is indicated if the samples show no trend in their means or variances. As an example, trace plots for the regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ for all 3 chains can be produced using the following code, and are shown in Figure 4.

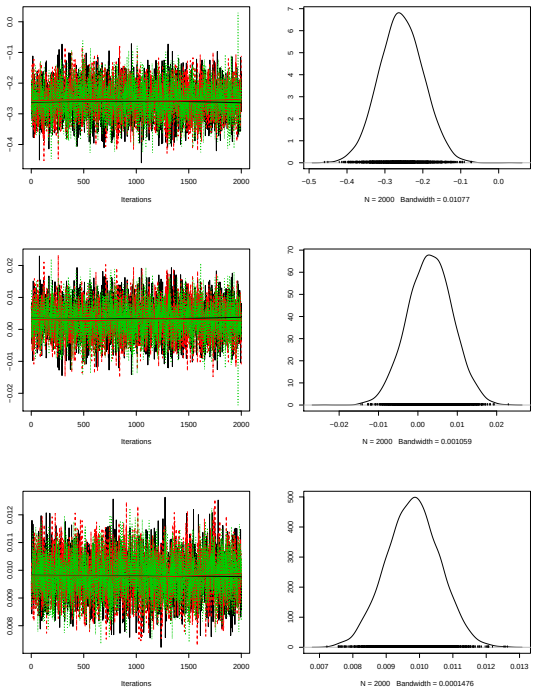

Figure 4: Traceplots of the MCMC samples from each chain.

```
library(coda)
beta.samples <- mcmc.list(chain1$samples$beta, chain2$samples$beta, chain3$samples$beta)
plot(beta.samples)
```

The figure shows the chains appear to have converged, as there is no change in the mean or variance of the samples between the three chains. An additional check is the Gelman-Rubin diagnostic (Gelman et al., 2013), which uses the between to within chain variation in the MCMC samples to quantify the potential scale reduction that might be achieved if we run the MCMC chains for longer. Gelman et al. (2013) recommend that a value less than 1.1 indicates good mixing of the chain, and the Gelman-Rubin statistic for each chain separately and then jointly for all chains is computed using the code below.

```
gelman.diag(beta.samples)
```

```
## Potential scale reduction factors:
##
```

```
##      Point est. Upper C.I.
## [1,]          1       1.01
## [2,]          1       1.01
## [3,]          1       1.00
##
## Multivariate psrf
##
## 1
```

The results shows that the samples appear to be well mixed, which again indicates that one can proceed with inference from this model. Note, in principle these diagnostic checks should be applied to every parameter, but due to the large number of random effects this is infeasible in practice. Therefore a pragmatic strategy is to undertake these checks for $(\boldsymbol{\beta}, \tau^2, \rho_S, \rho_T)$ and a sample of the spatio-temporal random effects $\{\psi_{kt}\}$. The results from each chain can be summarised by the `print()` function, which when applied to `chain1` gives the following output.

```
print(chain1)
```

```
##
## #################
## #### Model fitted
## #################
## Likelihood model - Poisson (log link function)
## Latent structure model - Autoregressive CAR model
## Regression equation - Y ~ offset(log(E)) + PM25 + IMD
##
## ############
## #### Results
## ############
## Posterior quantities for selected parameters and DIC
##
##              Median      2.5%       97.5%  n.effective Geweke.diag
## (Intercept)  -0.2601   -0.3746   -0.1409        1165.9        -1.2
## PM25          0.0033   -0.0082    0.0148        1154.0         0.9
```

```
## IMD           0.0098    0.0082    0.0114        1577.7         0.8
## tau2          0.0252    0.0226    0.0280        2000.0         0.7
## rho.S         0.9922    0.9855    0.9961        2000.0         1.1
## rho.T         0.8795    0.8586    0.8991        1858.5        -0.9
##
## DIC =  39345.94       p.d =  1869.179       LMPL =  -20003.09
```

The output provides a description of the model at the top, and model fit criteria such as the deviance information criterion (DIC, Spiegelhalter et al., 2002)) at the bottom. The middle table of resuts presents parameter summaries, including the posterior median point estimate (`Median`) and 95% credible intervals (`2.5%, 97.5%`). The `n.effective` column estimates the effective number of independent samples, and the `Geweke.diag` gives the Geweke diagnostic (Geweke, 1992), another MCMC convergence diagnostic that should lie between -2 and 2 to indicate convergence.

5. Inference from the model

We now answer the three questions of interest motivating the analysis posed in Section 2. Firstly, the effects of `IMD` and `PM25` on pneumonia mortality risk are quantified as relative risks, for a fixed increase ξ in each covariates value. For example, the relative risk for a ξ increase in `IMD` is computed by

$$\begin{aligned} RR(\texttt{IMD}, \xi) &= \frac{\text{Risk of disease if IMD increased by } \psi}{\text{Risk of disease given the current value of IMD}} \\ &= \frac{\exp(\beta_0 + \beta_1(\texttt{IMD}_k + \xi) + \beta_2 \texttt{PM25}_{kt} + \psi_{kt})}{\exp(\beta_0 + \beta_1 \texttt{IMD}_k + \beta_2 \texttt{PM25}_{kt} + \psi_{kt})} \\ &= \exp(\beta_1 \xi). \end{aligned} \tag{7}$$

Here we use the standard deviation of each covariate as the increase ξ, because they represent realistic increases in each covariates value. These increases are $2.26 \mu g m^{-3}$ for `PM25` and 8.02 for `IMD`. To compute these relative risks we first construct a matrix of the MCMC samples for the regression parameters (β_1, β_2) from all chains, and then compute the estimated relative risks as shown below.

```
beta.samples.combined <- rbind(chain1$samples$beta[ ,2:3], chain2$samples$beta[ ,2:3],
        chain3$samples$beta[ ,2:3])
round(quantile(exp(sd(dat.ordered$PM25) * beta.samples.combined[ ,1]), c(0.5, 0.025, 0.975)),3)
```

```
##   50%  2.5% 97.5%
## 1.007 0.982 1.033
```

```
round(quantile(exp(sd(dat.ordered$IMD) * beta.samples.combined[ ,2]), c(0.5, 0.025, 0.975)),3)
```

```
##   50%  2.5% 97.5%
## 1.082 1.068 1.095
```

The results show that the posterior median relative risk for `PM25` is close to 1, and that `PM25` is not significantly related to pneumonia mortality risk as the 95% credible interval contains the null risk of 1. In contrast, the index of multiple deprivation `IMD` is significantly related to pneumonia mortality risk, with a 95% credible interval that is wholly above 1. The posterior median relative risk is 1.082, which suggests that if `IMD` increases by 8.02 then the risk of pneumonia mortality increases by 8.2%.

The remaining questions of interest concern the spatio-temporal trends in disease risk $\{\theta_{kt}\}$, and the posterior risk distributions for each local authority and year are created by the code below, which divides the samples of fitted values $\{\mathbb{E}(Y_{kt}) = E_{kt}\theta_{kt}\}$ by the fixed expected numbers of disease cases $\{E_{kt}\}$.

```
fitted.samples.combined <- rbind(chain1$samples$fitted, chain2$samples$fitted,
        chain3$samples$fitted)
n.samples <- nrow(fitted.samples.combined)
n.all <- ncol(fitted.samples.combined)
risk.samples.combined <- fitted.samples.combined /
        matrix(rep(dat.ordered$E, n.samples), nrow=n.samples, ncol=n.all, byrow=TRUE)
```

Each column of `risk.samples.combined` contains the posterior samples for θ_{kt} for a single local authority and time period, and the columns are ordered in the same way as the rows of `data.ordered`. To estimate the average temporal trend to answer the second motivating

question, the first step is to estimate the average (mean) risk across the $K = 322$ local authorities for each year and MCMC sample, yielding the posterior distribution of these spatial averages for each year. This averaging is carried out by the code below.

```
N <- length(table(dat.ordered$Year))
risk.trends <- array(NA, c(n.samples, N))
  for(i in 1:n.samples)
  {
  risk.trends[i, ] <- tapply(risk.samples.combined[i, ], dat.ordered$Year, mean)
  }
```

Then the posterior median and 95% credible intervals can be computed for the spatially averaged risk as follows.

```
time.trends <- as.data.frame(t(apply(risk.trends, 2, quantile, c(0.5, 0.025, 0.975))))
time.trends <- time.trends %>% mutate(Year=names(table(dat.ordered$Year)))
colnames(time.trends)[1:3] <- c("Median","LCI", "UCI")
```

Then finally the estimated temporal trend in disease risk can be plotted as shown below, and is displayed in Figure 5.

```
ggplot(time.trends, aes(x = factor(Year), y = Median, group=1)) +
    geom_line(col="red") +
    geom_line(aes(x=factor(Year), y=LCI), col="red", lty=2) +
    geom_line(aes(x=factor(Year), y=UCI), col="red", lty=2) +
    scale_x_discrete(name = "Year", breaks=c(2002, 2005, 2008, 2011, 2014, 2017),
         labels=c("2002", "2005", "2008", "2011", "2014", "2017")) +
    scale_y_continuous(name = "Risk") +
    theme(text=element_text(size=16), plot.title=element_text(size=18, face="bold"))
```

The figure shows a clear downward trend in pneumonia mortality risk over the 16 year study period, which appears to flatten out from around 2014 onwards. We display the spatial pattern in disease risk in two ways, the first being the posterior median risk surface $\hat{\theta}_{kt}$. The second is the posterior exceedance probabilities (PEP), which are given by

$$\varphi_{kt} \;=\; \mathbb{P}(\theta_{kt} > 1|\mathbf{Y}), \tag{8}$$

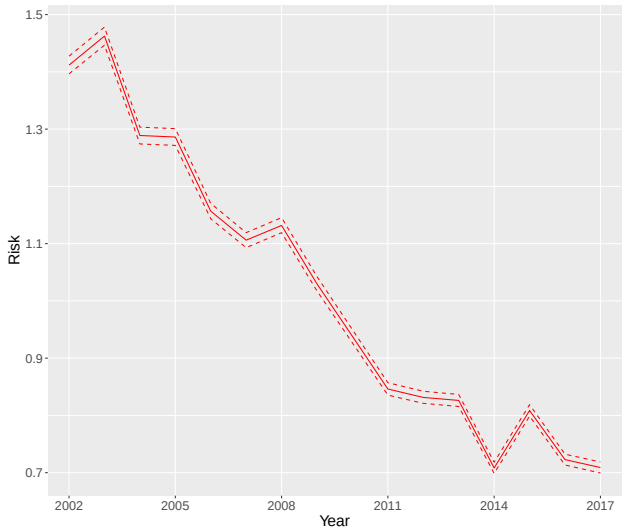

Figure 5: Posterior median and 95% credible interval for the temporal trend in disease risk.

the posterior probability that disease risk θ_{kt} is greater than one given the data $\mathbf{Y}$, where a risk of one is the average risk across England over the 16 year study duration. We illustrate the computation of both these quantities (median risk and PEP) for 2010 as follows:

```
risk.samples.2010 <- risk.samples.combined[ ,dat.ordered$Year==2010]
risk.2010 <- apply(risk.samples.2010, 2, median)
pep.2010 <- apply(risk.samples.2010 > 1, 2, mean)
```

Then, both quantities are added to the `SpatialPolygonsDataFrame` data set `residuals2010.LA` for mapping as follows.

```
residuals2010.LA$risk.2010 <- risk.2010
residuals2010.LA$pep.2010 <- pep.2010
```

The maps of the median risk and the PEP for 2010 are displayed in Figure 6, and were generated with similar code to that used to create Figure 3 which is hence not shown for brevity. The posterior median risk map (left) shows that the risks are highest in the north-west of England close to the cities of Liverpool, Manchester and Sheffield, while low-risk areas are generally the larger and more rural local authorities. The PEP map (right) shows that a substantial number of rural local authorities have a zero probability of exceeding a

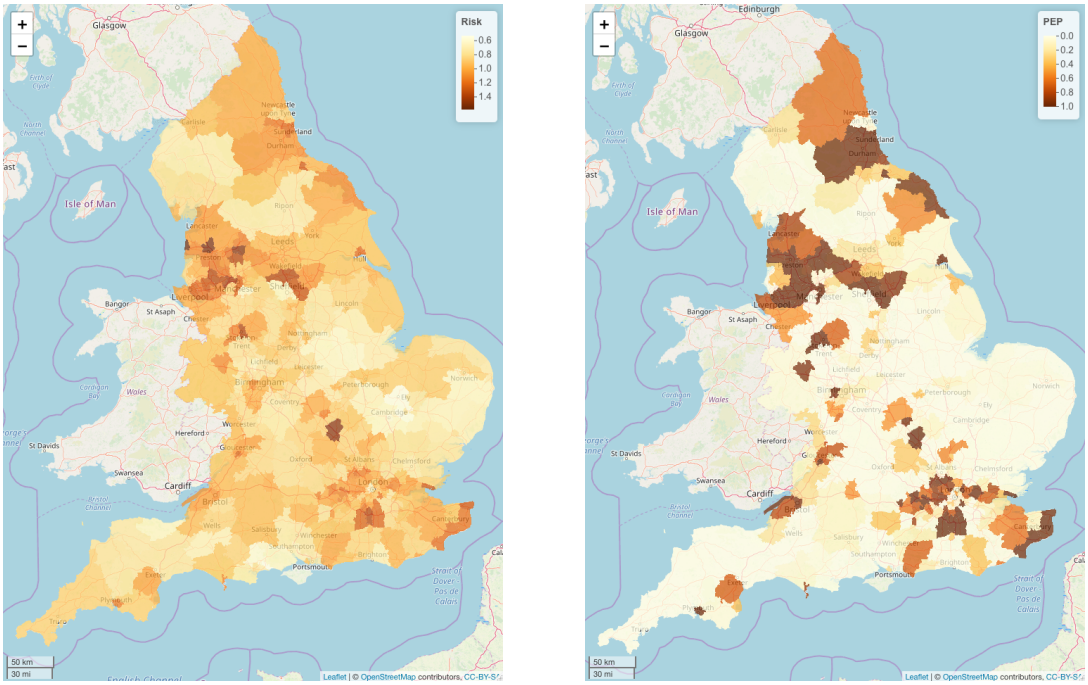

Figure 6: Estimated (posterior median) risk surface for 2010 (left) and the posterior exceedance probabilities that that risk in 2010 is greater than 1 (right).

risk of one, while the probabilities for the urban areas, particularly in the north are much greater.

Finally, the third motivating question concerns health inequalities, and here we focus on *total inequality* (World Health Organisation, 2013), which measures the variation in disease risk over the study region. We quantify this variation by the interquartile range (IQR) separately for each year, which measures the difference between the third and first quartiles of disease risks for a given year. These IQRs are computed for each year as shown below.

```
risk.median <- apply(risk.samples.combined, 2, median)
inequality <- tapply(risk.median, dat.ordered$Year, IQR)
```

The temporal trend in these IQRs are displayed in Figure 7, and the code for creating this figure is not shown as it is similar to that used to create Figure 5. The figure shows that total inequality in pneumonia mortality risk, as measured by the IQR, has reduced by around a half, with values around 0.3 in 2002 compared to around 0.16 in 2017. This suggests that health inequalities are narrowing over time, suggesting that the population is becoming more

Figure 7: Estimated temporal trend in the health inequality in pneumonia mortality risk as measured by the spatial interquartile range.

even in terms of risk in later years.

6. Conclusions

In this tutorial we have illustrated how to analyse spatio-temporal areal unit data using the `CARBayesST` package in `R`, which fits models in a Bayesian setting via MCMC simulation. The worked example on pneumonia mortality in England describes a complete spatio-temporal analysis, beginning with data input, wrangling and visualisation, and then modelling and drawing inference from the fitted model to answer epidemiologically important questions. The package allows the estimation of disease risk trends in both space and time, and the sister package `CARBayes` (Lee, 2013) allows purely spatial data to be modelled in a similar way.

There are two main areas for future work in this software field, the first being the extension of the class of models and data structures that can be handled by these packages. While `CARBayes` has limited capability for modelling multivariate disease data, there is currently no capability for modelling multivariate areal unit data structured in space and time. The development of such models is a research priority due to the increasing availability of data on multiple diseases at the same spatio-temporal resolution. The second area for future development is the creation of an easy-to-use point-and-click user interface for the `CARBayes` /

`CARBayesST` family of software, because as the worked example in this tutorial demonstrates, modelling of such data requires a certain level of knowledge in `R` which may not be available to all applied researchers.

References

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., Songini, M., 1995. Bayesian analysis of space-time variation in disease risk. Statistics in Medicine 14, 2433–2443.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration with two applications in spatial statistics. Annals of the Institute of Statistics and Mathematics 43, 1–59.

Bivand, R., Pebesma, E., Gomez-Rubio, V., 2013. Applied Spatial Data Analysis with R. Springer-Verlag.

Blangiardo, M., Cameletti, M., Baio, G., Rue, H., 2013. Spatial and spatio-temporal models with R-INLA. Spatial and Spatio-temporal Epidemiology 4, 33–49.

Department for Communities and Local Government, 2015. The English Indices of Deprivation 2015.

Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., Beard, J., 2007. Evaluating the effect of neighbourhood weight matrices on smoothing properties of conditional autoregressive (CAR) models. International Journal of Health Geographics 6, 54.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2013. Bayesian Data Analysis. 3rd ed., Chapman and Hall / CRC.

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, in: Bayesian Statistics, University Press. pp. 169–193.

Illian, J., Sørbye, S., Rue, H., 2012. A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). Annals of Applied Statistics 6, 1499–1530.

Knorr-Held, L., 2000. Bayesian modelling of inseparable space-time variation in disease risk. Statistics in Medicine 19, 2555–2567.

Lawson, A., 2018. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, Third Edition. Chapman and Hall/CRC.

Lawson, A., Lee, D., 2017. Bayesian Disease Mapping for Public Health. Elsevier. chapter Handbook of statistics 36: Disease Modelling and Public Health, Part A, Halloran, A. Rao, S. Pyne and C. Rao (eds). pp. 443–481.

Lee, D., 2013. CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. Journal of Statistical Software 55, 1–24.

Lee, D., Rushworth, A., Napier, G., 2018. Spatio-temporal areal unit modeling in R with conditional autoregressive priors using the CARBayesST package. Journal of Statistical Software, Articles 84, 1–39.

Leroux, B., Lei, X., Breslow, N., 2000. Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence. Springer-Verlag, New York. chapter Statistical Models in Epidemiology, the Environment and Clinical Trials, Halloran, M and Berry, D (eds). pp. 135–178.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B 73, 423–498.

Lunn, D., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing 10, 325–337.

Martins, T., Simpson, D., Lindgren, F., Rue, H., 2013. Bayesian computing with INLA: New features. Computational Statistics and Data Analysis 67, 68–83.

Moraga, P., 2018. Small Area Disease Risk Estimation and Visualization Using R. The R Journal 10, 495–506.

Moran, P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S., Gelman, A., DiMaggio, C., 2019. Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in STAN. Spatial and Spatio-temporal Epidemiology 31, 100301.

Robert, C., Casella, G., 2010. Introducing Monte Carlo Methods with R. Springer.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested laplace approximations (with discussion). Journal of the Royal Statistical Society Series B 71, 319–392.

Rushworth, A., Lee, D., Mitchell, R., 2014. A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. Spatial and Spatio-temporal Epidemiology 10, 29–38.

Rushworth, A., Lee, D., Sarran, C., 2017. An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk. Journal of the Royal Statistical Society Series C 66, 141–157.

Spiegelhalter, D., Best, N., Carlin, B., Van der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B 64, 583–639.

Womble, W., 1951. Differential systematics. Science 114, 315–322.

World Health Organisation, 2013. Health Inequality Monitoring with a special focus on low- and middle-income countries.